NOT TO BE FURTHER COPIED OR REPRODUCED

(* BATCHES *)

5

Visual Summaries
(schematic plots)

		Exh.	Page
Introduction	5-1	1 a	5-5
		2	5-8
5A. Graphical display	5-2		
- dot plots,	5-4 to 5-7	3 b	5-13
- stem-and-leaf compared,	5-7	4	5-15
- using tracing paper,	5-7 to 5-9		
- general remarks,	5-10		
		5 c	5-18
5B. Box plots	5-10		
- box-and-dot combined,	5-14	6 d	5-20
		7	5-23/25
5C. Learning to do better	5-14		
outside observations,	defined at 5-17		
side values,	defined at 5-17	8 e	5-27
dh, as a mark,	defined at 5-17	9	5-29
three-halves values,	defined at 5-17	10	5-31
detached observations,	defined at 5-17		
		11 f	5-34
5D. Schematic plots	5-19	12	5-36
- how many are reasonable?,	5-21	13	5-37
- details of calculation,	5-22 to 5-26	14	5-38
		15	5-40
5E. Pros and cons; the Rayleigh example	5-26	16	5-41
- easing the task,	5-28 to 5-30		
- aim reiterated,	5-30 to 5-32		
- choice of plots,	5-32 to 5-33	17 g	5-45
		18	5-46
5F. Some further examples	5-33	19	5-47
- ultimate hydroelectric powers,	5-33 to 5-34	20	5-48
- land areas of counties,	5-35	21	5-49
- Tippett's warp breaks,	5-35 to 5-39	22	5-50
- the march of temperature,	5-39 to 5-42		
		h	
5G. Different aims, different schematic plots	5-43	23 p	5-55
* * *			
5H. How far have we come?	5-50		
* * *			
5P. Problems for chapter 5	5-53		

5

Visual Summaries
(Schematic Plots)

Now that we have thought about pulling batches of numbers apart, it is time to turn again to pulling them together, to conveying -- either to people or to further analysis -- their general character.

So far as conveying general character to people goes, we have stem-and-leaf pictures as a leading contender. Anything we consider for this purpose ought to do better in at least one important way.

Conveying general character to further analysis is quite a different matter. To do this effectively, we are likely to have to use numbers, to be both specific and quantitative in our summaries. We shall turn to this in later chapters.

Throughout, we will need to keep in mind the possibility -- and the frequent need -- of re-expressing our basic numbers before we summarize them. In working with stem-and-leaf pictures we identified at least three goals to be approached by re-expression:

- elimination of unnecessary apparent strays.
- increase in symmetry (often very closely tied to reduction of strays).
- approach to equal spread when two or more batches are to be compared.

Notice that moving toward any one of these goals tends to make summarization easier and more effective. If the pattern of the batch is symmetric and free of strays we can describe it more simply. Perhaps we need only say -- or show -- where it is centered and how widely it is spread. Similarly, if we can make the spreads of the different batches nearly the same, one number -- or one impression -- may do well enough to describe the spread of all. Moreover, and much more importantly, similar spread facilitates comparison.

We must expect to continue to press on both fronts:

- finding better ways to summarize.
- making the data easier to summarize.

5A. Graphical display

One natural and insightful reaction to much of the discussion of chapter 4 is:

- Why waste time with arithmetic -- graphs would have done all that quicker and easier.

Much can be said for such a view, including:

- if those things attempted in chapter 4 made up all there was of data analysis, graphs and no arithmetic would indeed be easier.
- when they can be used, graphs do convey more to people than does arithmetic -- more in impact and more in diversity.

But conversely, there are important things to be said in the opposite direction:

- there is no substitute for the combination of graphs and arithmetic.

- there are many problems of data analysis where much arithmetic has to be used to break things out before graphs can show them to us. (In particular, numbers are much more susceptible to further analysis than are graphs.)

- much of what we were carefully explicit about in discussing our arithmetic in chapter 4 still occurs, perhaps tacitly or in the mind's eye, with graphs.

In view of all this, we can recognize two important reasons why we took up arithmetic when we did:

1) we will need generalizations and analogs of the simple arithmetic of chapter 4 as soon as we come to look at data of even slightly more complex structure.

2) taking arithmetic first forces us to concentrate on how it operates.

Here in chapter 5 we shall begin to get graphing of batches of numbers rolling with equal vigor, so that we can use graphs and arithmetic together, each for what it does best.

We do not wish to overstress arithmetic. Arithmetic is a powerful tool, but one that, like all tools, takes effort. It is needed to do many things. Because a modern computer can do much arithmetic cheaply and easily, we can be very tolerant of having a lot of it done for us. (Especially in hand calculation, however, we still want to go as far as we can for a little effort.)

Arithmetic, wisely used, can do much to break through into the significant areas of our data. If the facts are

simple enough, it may even manage to communicate these facts to us. In general, however, we must look to pictures as the most effective way of communicating to people quantitative information that is at all complex. There is usually no substitute for the friendly graph.

* dot plots *

Let us return to the ten bowling scores of chapter 4. Exhibit 1 shows them in two versions. On the left we have plotted them in a box which starts at zero, on the right in a magnified box that starts at 182.

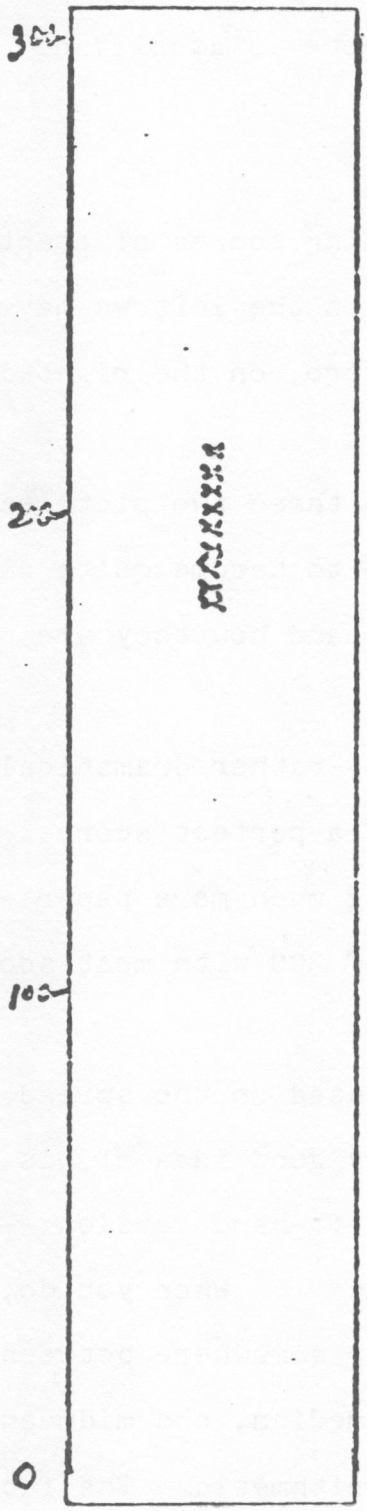
The most obvious fact is that these two pictures serve very different purposes. We need to become quite clear as to what these different purposes are and how they are served by different displays.

The left-hand version shows -- rather dramatically -- that our bowler is getting about 2/3 of a perfect score. It is saying -- with more impact, but using much more paper -- this bowler is "scoring about 200 out of 300 with most scores reasonably close to 200".

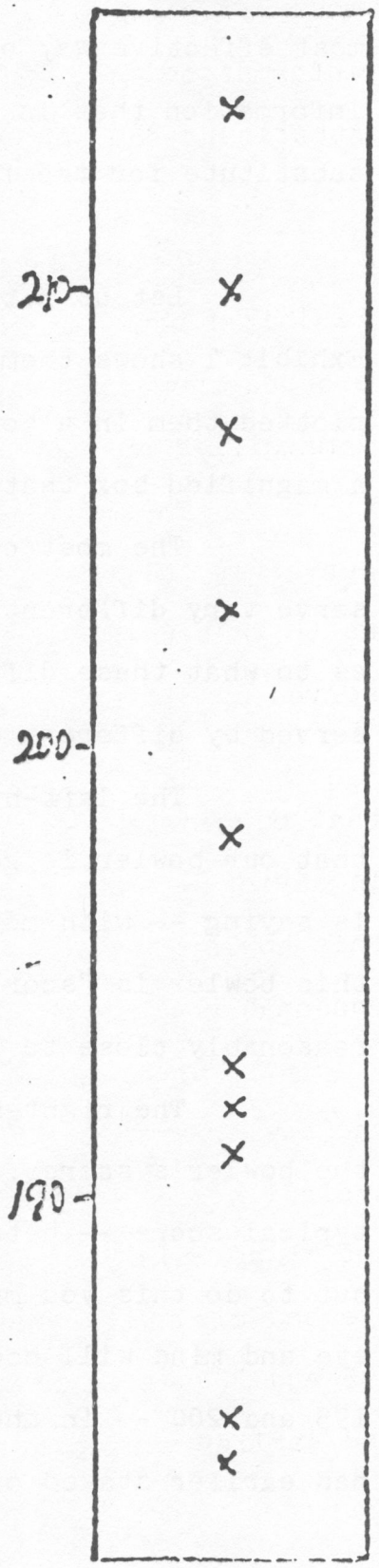
The right-hand version is focused on the spread of the bowler's scores. You can get a very good idea of his typical score -- better than from the left-hand version -- but to do this you must read the scale . When you do, eye and mind will cooperate on an answer somewhere between 195 and 200 -- in the range that mean, median, and midmean had earlier staked out as a matter of arithmetic. The facts

Two representations of the 10 bowling scores

A) IN THE LARGE



B) IN THE SMALL



about central value are there for anyone who wants access to them, but what strikes the eye most heavily in the right-hand version is information about the amount and pattern of spread of the scores. We are quite likely, for instance, to notice that higher scores seem more spread out from one another than do lower scores -- to do this even if we had not come to our picture with such a possibility in mind.

One of the great features of graphical impact -- one of the reasons which arithmetic can never serve ALL the functions of graphs -- is the graph's ability to show you what you were not looking for -- what you may even have been very reluctant to see.

In chapter 4 we talked about separating our data in the simplest way -- into central values and residuals. To do this by arithmetic is routine, detailed, sometimes tedious (unless an automatic computer can do it for us in seconds). We have really done much the same thing in exhibit 1: the left-hand version tells us about where the batch is centered, the right-hand version tells us about residuals. Separation is still a major concern, though with graphs it is easier to fail to notice that it is happening. This is particularly so since the right-hand picture tells us almost all about residuals without having chosen a particular central value. Arithmetic would have to choose such a value, a value that would then be subtracted to make explicit residuals. The

residuals in the right-hand picture are implicit, flexibly defined, provided by a mechanism easy to overlook. Yet their message pounds in on us whether or not we like it.

(* stem-and-leaf compared *)

Stem-and-leaf displays go a long way toward graphs, both in nature and in impact. In fact, anything worthy of the name display has to be a substantial part of the way from pure numbers toward a graph. There are many senses in which a stem-and-leaf picture is a poor man's graph, where poverty may well be merely a lack of available time or effort, an inability to do as much as to plot a graph.

Let us look at a stem-and-leaf for the 10 bowling scores as a reminder and a comparison. Exhibit 2 shows us the possibilities. Rather clearly, panel A of exhibit 1 showed us the general size better than panel A of exhibit 2.

So far as spread and pattern go, however, panel A of exhibit 1 is far inferior to the other three, for which effectiveness increases in the order:

- panel A of exhibit 2 (raw stem-and-leaf).
- panel B of exhibit 2 (stem-and-leaf for residuals).
- panel B of exhibit 1 (magnified dot plot).

The last is best, but not by as much as we might have expected.

(* using tracing paper *)

We want our graphs easily made and of strong impact. At first sight, these might seem to be contradictory requirements. They need not be if we are both wise and careful

exhibit 2 of chapter 5

Stem-and-leaf for the 10 bowling scores

A) RAW SCORES

22	
21	04
20	37
19	1238
18	45
17	

B) RESIDUALS from the TRIMEAN of 197

1	7
1	03
0	6
0	1
-0	4
-0	65
-1	32
-1	

To graph well we need two things:

- graph paper (warnings to come in chapter 8).
- tracing paper (preferably padded).

Just how we use them depends on how much we want to save money.

If we are willing to expend one sheet of each on a plot, we make the plot first on graph paper and then trace it on tracing paper.

If we feel more economical, we slip the graph paper under the tracing paper and use the "show through" of the graph paper lines in making the plot on the tracing paper.

There seems to be a natural complementarity between graph paper and tracing paper:

- while making a graph we can make good use of both the finely spaced lines on the graph paper and up to a dozen or so numbered ticks on each axis.
- in looking at a graph to be shown what goes on (not as a substituted for a table of numbers) we can make good use of both the total absence of lines across the plot and of no more than three (perhaps 4) numbered ticks on each axis (a few unnumbered ticks can often be added).

To make a graph, details are important. To use it they must not be allowed to distract our attention. (If details are important, we should be looking at a different graph, where the important things are no longer details.)

(* general remarks *)

Choosing which graph we plot or look at is very important. Most books or articles on graphical presentation assume that our task resembles showing the obvious to thick-headed viewers. Accordingly they emphasize getting across where batches are centered visually and have much to say about the importance of starting scales from zero. If the graph is needed to convey that 200 is about 2/3 of 300, this is exactly the right view. (For those newspaper readers least familiar with numbers, this may be exactly the challenge.) But if the graph is to meet other needs, it needs to be otherwise constructed.

We are in the detective business. The picture that shows the big desk, the oriental rug on the floor, and the three windows can help the investigator of crime, but not enough. The close-up views of the match on the rug, the scratch on the desk, or the fingerprint on the third window are at least equally important. We will not shun the distant-view graph in our data analysis, but we will plan to make special use of enlarging graphs that enlarge things until we see them clearly.

5B. Box plots

To see the 10 individual bowling scores on a graph can be helpful, but even with so few values we tend to find detail interfering with a general view. What if we have a batch of 218? The time needed to plot all 218 points should not be

taken; even we could afford it, the result not be easy to look at. Stem-and-leaf arrays will be better than dot plots for large batches.

If we are to do better with plots than with stem-and-leaf arrays for large batches, we must do something more. We must eliminate detail that is usually less valuable, thus focusing our attention on what are usually -- though not always -- the more important aspects of the batch. We must make visual summaries -- we must make schematic plots.

Which of the characteristics of a batch are we reasonably sure should be shown to us? Surely the elements of the 5-number summary:

- the median.
- the hinges.
- the extremes.

Are these enough? Almost, but not really. They do tell us a lot, but as we shall see shortly, a little more can often be a significant help.

For the present, however, let us carry on with only these five values, choosing a pattern to which we can add further information. To show hinges, median and extremes, we shall adhere to this basic pattern:

- a closed box with ends at the hinges, thus covering the middle of the batch.
- extremes shown by some form of dot.
- a bar across the box for the median.

If a middle mean is wanted, it can be found easily either mentally or by pointing. We merely:

- first, spot a point halfway between the hinges.
- then, spot a point halfway between the median and this first point.

This locates the trimean, easily, rapidly, and -- after a little practice -- with surprising accuracy.

The measures of spread that we have already noticed as accompanying the five-value summary are clearly also before us:

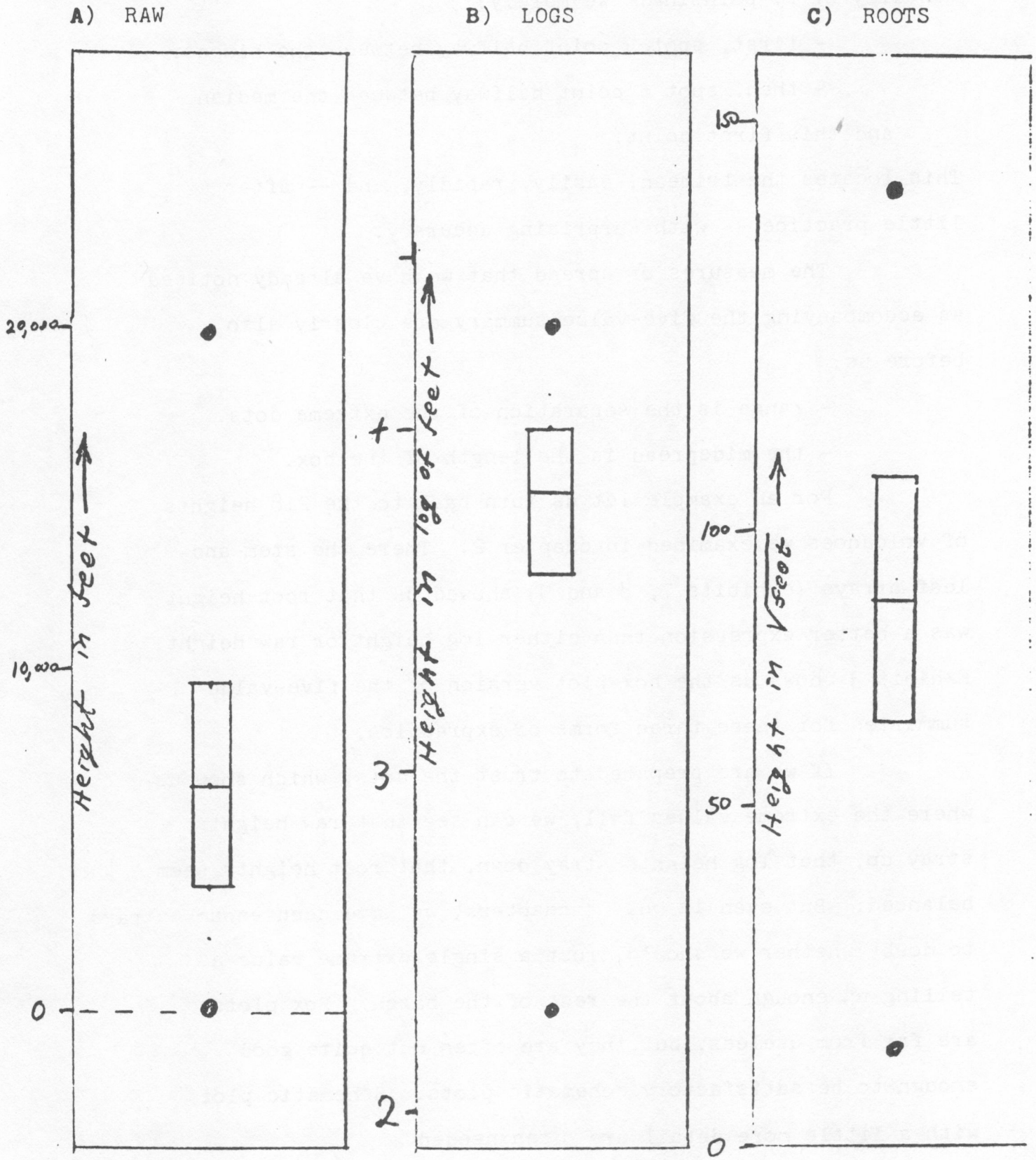
- range is the separation of the extreme dots.
- the midspread is the length of the box.

For an example let us turn back to the 218 heights of volcanoes we examined in chapter 2. There the stem-and-leaf arrays (exhibits 7, 8 and 9) showed us that root height was a better expression than either log height or raw height. Exhibit 3 shows us the box-plot version of the five-value summaries for these three forms of expression.

If we are prepared to trust the dots, which show us where the extreme values fall, we can see that raw heights stray up, that log heights stray down, that root heights seem balanced. But even in only 4 chapters, we have seen enough strays to doubt whether we should trust a single extreme value as telling us enough about the rest of the batch. Box plots are far from useless, but they are often not quite good enough to be satisfactory schematic plots. Schematic plots with a little more detail are often needed.

Height of 218 volcanoes shown by box plots

using three different expressions



(* box-and-dot combined *)

Now that we have both dot plots and box plots at hand, we ought to ask how to combine them. Exhibit 1, above, showed the two extreme purposes -- central value and residuals -- approached through the dot plot. Exhibit 4 shows the same two extreme purposes approached through a suitable combination of dot plots and box plots. Panel A of this exhibit shows a box plot, alone, concentrating our attention on the central value as hard as we know how without losing some feeling of spread.

Panel B, on the other hand, combines a dot plot with a box plot, the visual strength of the latter being weakened enough -- by the use of broken lines -- to give the dots their fair share of our attention. Proper adjustment of attention can be special in many displays; the doubting reader is urged to copy panel B with firm solid lines and look at it. Each of these two portrayals goes a long way toward combining a very effective approach to both aims.

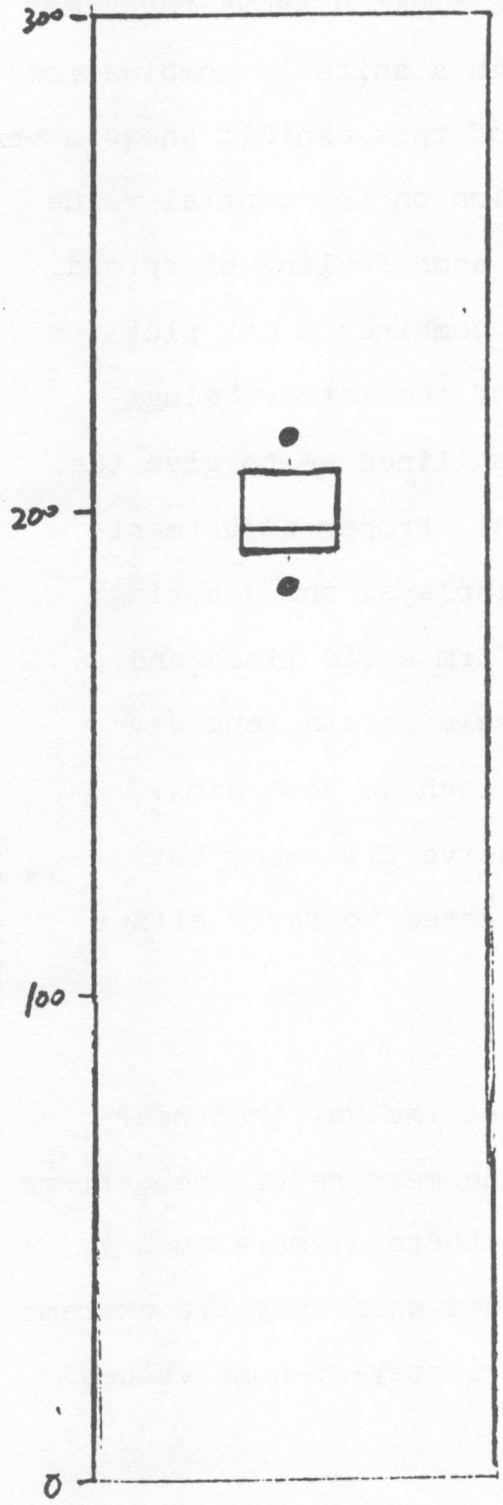
There is no one picture to serve all aims, but there can be effective compromises, slanted to favor either extreme.

5C. Learning to do better

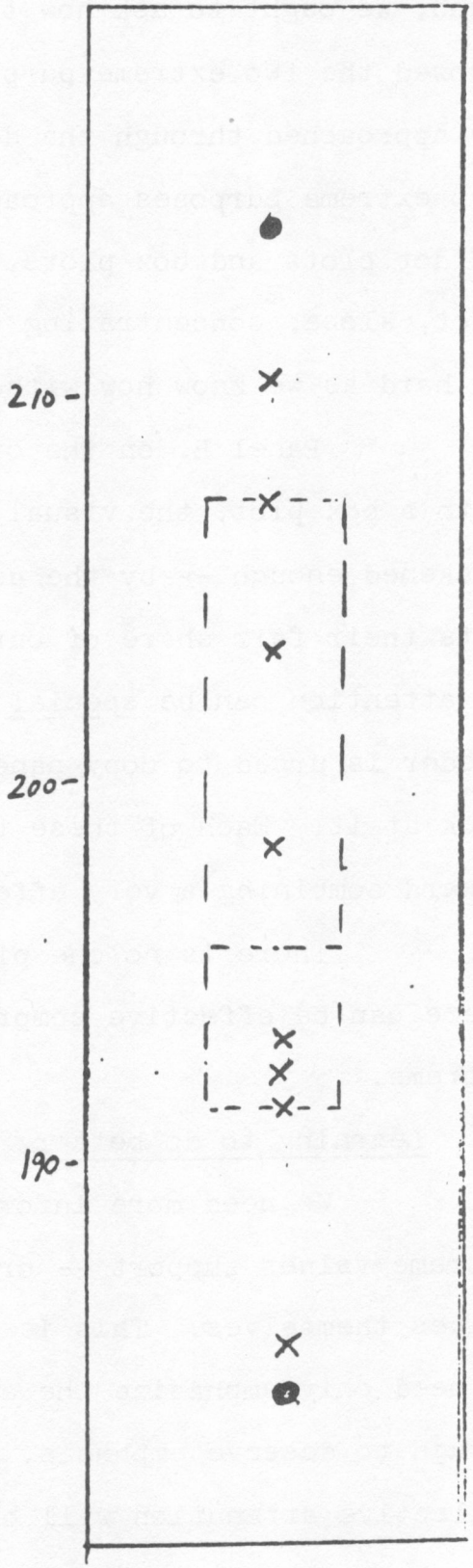
We need more information about the way that near-extreme values support -- or deny -- the message of the extreme values themselves. This is clear, but there is more to say. We need only emphasize the extreme values when they are extreme enough to deserve emphasis. If they are stay-at-home values, collective attenution will be enough.

Box-and-dot combined for the 10 bowling scores

A) In the LARGE



B) In the SMALL



This philosophy clearly calls for some sort of cutting points to divide the stay-at-homes from the rovers.

Except in Scotland, where "not proven" is a legal verdict, processes of criminal judgement are expected to reach "guilty" or "not guilty" verdicts. The need for such sharpness in a legal system is clear.

When we come to confirmatory data analysis, we can not only allow "scotch verdicts", we must face the need for more refined gradations. In exploratory data analysis, we ought to regard avoiding "yes" or "no" splits as one of our prime requisites.

Given a batch of values, what easy technique can we use to identify some -- possibly none -- of the values as far enough from the bulk of the observations to be worth either special attention or mental detachment?

If it is to apply to any batch of values, our technique must compare individual rather extreme values with the "general run" of values in that same batch. In identifying this "general run" we need, by definition, to be resistant. In doing this, we must give some attention to level -- and some to spread. Thus we must start from at least 2 points.

The simplest resistant starting points are the hinges -- and their difference, the midspread. A little experimentation shows us that we do quite well by laying off one midspread beyond each hinge, and using the far ends as the inner dividing values. Since we find it convenient to call values beyond the inner dividing points

outside observations

it is convenient to call the inner dividing points

side values .

Some outside values are far out. Others may well be only slightly wandering stay-at-homes. We need, at least, to try to separate outside values into two classes.

For this we need outer dividing values. We could lay-off the same distance

midspread = difference of hinges

which we will mark "dh" to avoid too many marks beginning with "M", beyond the side values. Surely anything further out than this looks like a rover. If we try such a rule for a while, we find that it seems too stringent. So we split the difference and set

three-halves values

marked "T" just

$$\frac{3}{2}\,(dh)$$

outside the hinges and call all observations beyond them

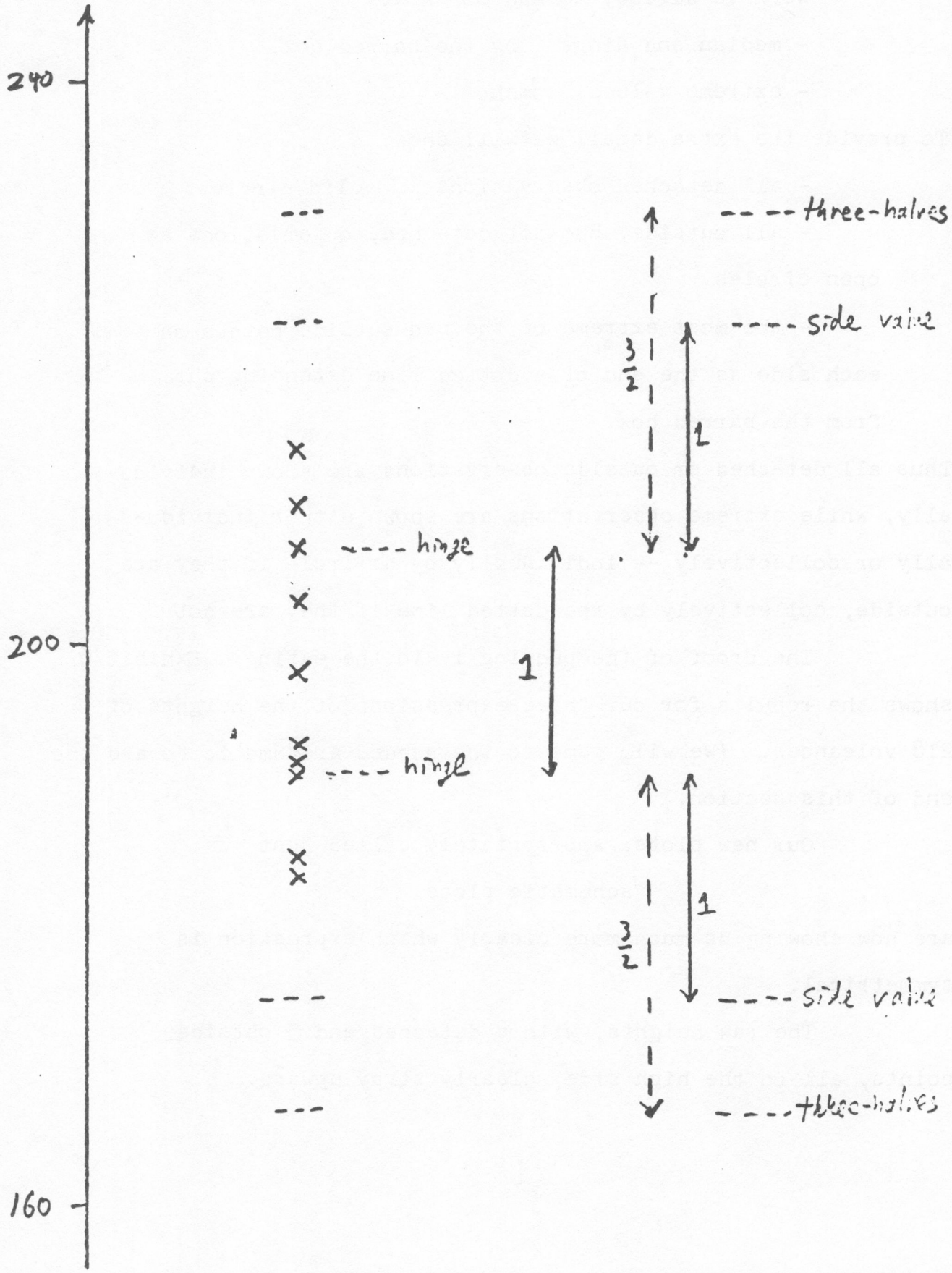
detached

Exhibit 5 shows how all four values lie with respect to a familiar batch, the 10 bowling scores.

With these new tools, we are ready to make better schematic plots.

Side values and three-halves values for the 10 bowling scores

(numbers on arrows show relative lengths)



5D. Schematic plots

We have already agreed to show:

- median and hinges, by the barred box.
- extreme values, somehow.

To provide the extra detail we will show:

- all detached observations as solid circles.
- all outside, but not detached, observations as open circles.
- the most extreme of the non-outside points on each side as the end of a dotted line extending out from the barred box.

Thus all detached or outside observations are shown individually, while extreme observations are shown either individually or collectively -- individually by a circle if they are outside, collectively by the dotted line if they are not.

The proof of the pudding is in the eating. Exhibit 6 shows the results for our three expressions of the heights of 218 volcanoes. (We will come to the needed arithmetic toward the end of this section.)

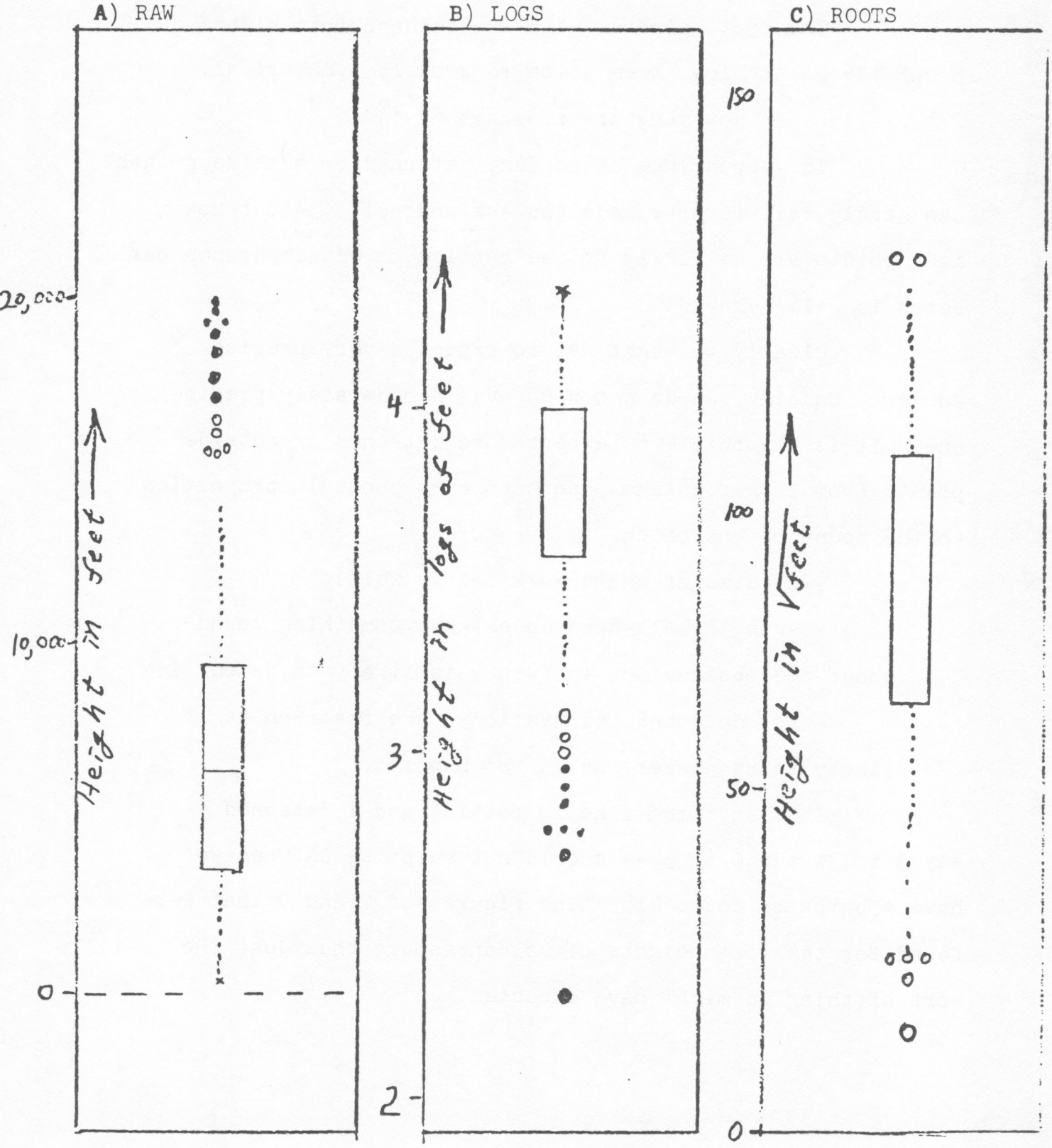
Our new plots, appropriately called just

schematic plots

are now showing us much more clearly which expression is symmetrical.

The raw heights, with 8 detached and 5 outside points, all on the high side, clearly stray upward.

Height of 218 volcanoes using three different expressions
(Schematic Plots)



The log heights, with 8 detached and 3 outside points, all on the low side, clearly stray downward.

The root heights, with 2 outside points high and 5 outside points low, seem quite reasonably symmetrical.

(* How many are reasonable? *)

The appearance of so many detached or outside points can hardly fail to have made you ask yourself: About how many points are we likely to see outside or detached when our batch is well-behaved?

Clearly we ought not to expect a very precise answer. Equally, we do not need even a moderately precise one. It is reasonable to expect more detached or outside points from larger batches, perhaps even about in proportion to the count of the batch.

Two rules of thumb work fairly well:

- even in well-behaved batches something round about one observation in twenty is likely to be outside.
- for detached observations, the fraction is likely to be nearer one in two hundred.

Thus anything like 10 outside and 1 detached -- say 3 to 30 and 0 to 5 -- should not surprise us when we have a batch of count 218. The figures of 7 and 0 that were found for the root heights of volcanoes are thus just the sort of thing we might have expected.

(* details of calculation *)

Exhibit 7 gives details for the three expressions of volcano heights. The stem-and-leaf, and the scrawl below it giving count, hinges and median, we have seen before. The new feature is the scrawl marked

```
dh|
S/T|
S/T|
out|
out|
```

Against dh this shows dh and, after "up", $\frac{1}{2}$ dh, where dh is, as noted above, the difference of hinges. Against each S/T are given a side value and a three-halves value, the side value lying dh outward of the corresponding hinge and the three-halves value lying $\frac{1}{2}$ dh outward of the side value. Against each out is shown, in parentheses, first the number of observations beyond the side value and then, after a comma, the number beyond the three-halves value.

Panel B shows three alternative forms of the scrawl of panel A, summarizing the raw heights. The first adds a line to note the adjacent observations, those nearest to being outside while not being so, listing them after adj. The second lists the values of the outside and detached observations after out, instead of giving only counts, separating the detached observations by a / . The third lists the outside and detached observations separately.

All three alternatives take somewhat more space. On the other hand, however, each of them gives every bit of detail needed to construct the corresponding symbolic plot:

Finding outside and detached observations

(see exhibits 7, 8, 9 of chapter 3)

A) THE RAW DATA IN FULL

	8	0 | 25666789
	16	1 | 01367799
	39	2 | 000112224445566677889 99
	57	3 | 01122445555667899
to	79	4 | 01112333334446788999 99
✓	23	5 | 00112223445566666677799
to	116	6 | 00114455666677788 9
	98	7 | 00001112334555678889
	78	8 | 122223335679
	66	9 | 000123344556779
	51	10 | 0112233445689
	38	11 | 0112334669
	28	12 | 11244456
	20	13 | 03478
	15	14 | 00
	13	15 | 667
	10	16 | 25
	8	17 | 29
	6	18 | 5
	5	19 | 03379

218	
H	38
M	65
H	96

dh	58 up 29
S/T	-20 / -49
S/T	154 / 183
out	(0,0)
out	(13,6)

Examples:

$$58 = 96 - 3$$

$$29 = \frac{1}{2}(58)$$

$$-20 = 38 - 58$$

$$-49 = -20 - 29$$

$$154 = 96 + 58$$

$$183 = 154 + 29$$

(Cont'd)

exhibit 7 of chapter 5 (Cont'd)

B) ALTERNATIVE SCRAWLS -- all for data of Panel A

dh	58 up 29
S/T	-20 / -49
S/T	154 / 183
out	none
out	(13,6)
adj	2,140

dh	58 up 29
S/T	-20 / -49
S/T	+154 / +183
out	none
out	156,156,157,162,165,172,179,/185,190,193,193,197,199

dh	58 up 29
S/T	-20 / -49
S/T	+154 / +183
out	none
det	none
out	156,156,157,162,165,172,179
det	185,190,193,193,197,199
adj	2,140

C) CONDENSED RESULTS for LOG and ROOT DATA -- see exhibits 8 and 9 of chapter 3 for stem-and-leaf displays

log heights

218	
H	158
M	181
H	198

dh	40 up 20
S/T	118/98
S/T	238/258
out	100,104,111
det	30,70,78,78,78,85,90,95
out	none
det	none
adj	120,230

(Cont'd)

exhibit 7 of chapter 5 (Cont'd)

root heights

218	
H	62
M	80
H	98
dh	36 up 18
S/T	26/8
S/T	134/152
out	14, 22, 25, 25, 25,
det	none
out	136,136
det	none
adj	26,132

- which points are to be plotted separately. (All listed.)

- which are to be filled in. (Beyond T.)

- where the dotted lines are to end. (Adjacent observations.)

Panel **C** gives the condensed results for log heights and root heights. The third alternative form of the scrawl is used for convenience.

We will feel free to use whichever form of the scrawl seems to fit in easily in a particular situation.

5E. Pros and cons; the Rayleigh example

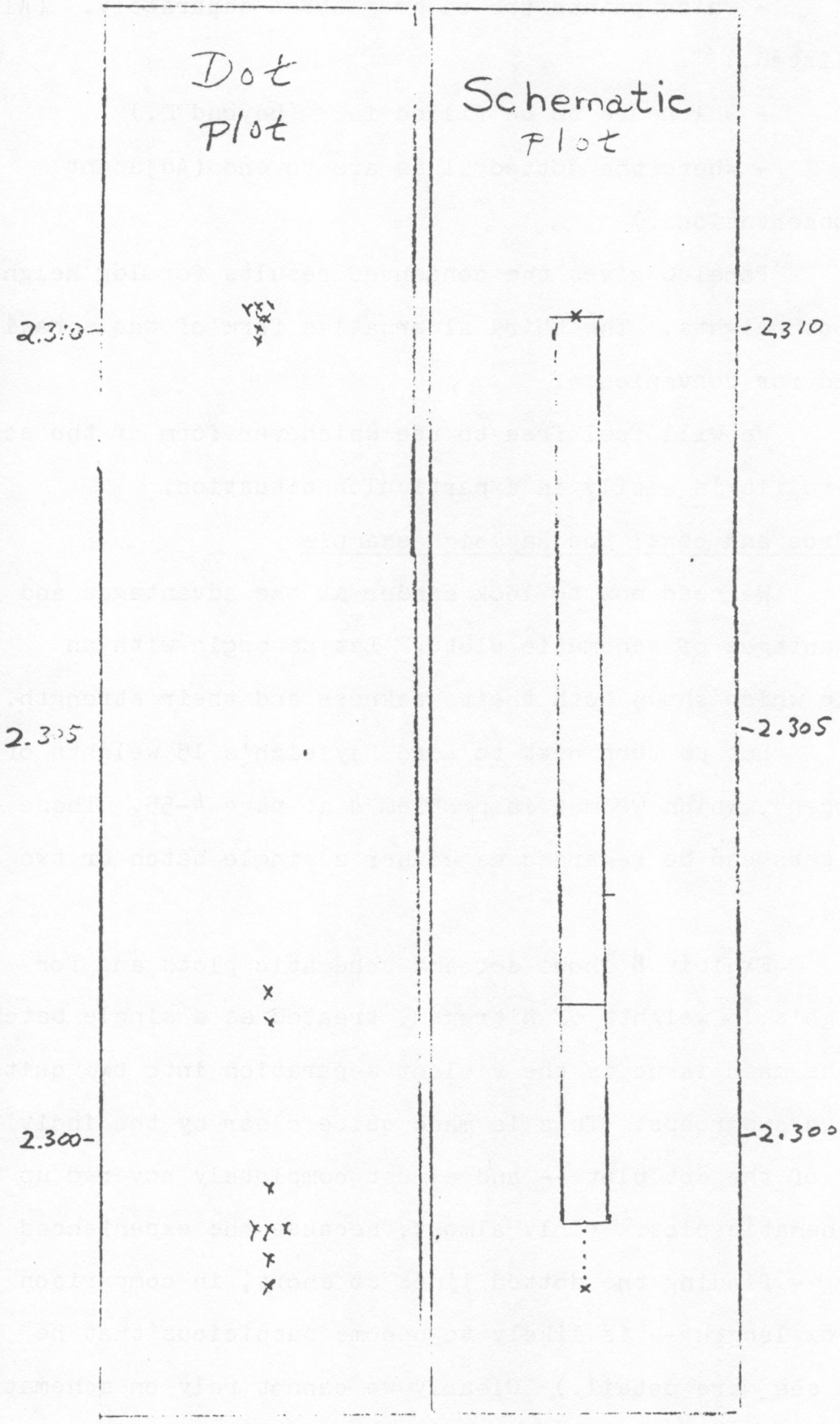
We need now to look harder at the advantages and disadvantages of schematic plots. Let us begin with an example which shows both their weakness and their strength.

Let us turn next to Lord Rayleigh's 15 weights of "nitrogen", which we met in problem 4 at page 4-55. These 15 weights can be regarded as either a single batch or two batches.

Exhibit **8** shows dot and schematic plots and for Rayleigh's 15 weights of nitrogen, treated as a single batch. Here the main issue is the violent separation into two quite isolated subgroups. This is made quite clear by the individual values of the dot plot -- and almost completely covered up by the schematic plot. (Only almost, because the experienced viewer -- finding the dotted lines so short, in comparison with box length -- is likely to become suspicious that he should see more detail.) Clearly we cannot rely on schematic

exhibit 8 of chapter 5

Rayleigh's 15 weights of "Nitrogen"



plots to call our attention to structure near the center of the batch, we must use some other approach when this is needed. (Either dot plots or stem-and-leaf displays will serve.)

Exhibit 9 uses the schematic plots for one of the purposes for which they are best fitted: comparison of two or more batches. In it the two batches of Rayleigh's weights, one for "nitrogen" from air, the other for "nitrogen" from other sources, are set out and compared. Clear to every eye are:

- the relatively huge difference between the batches.
- a slight tendency of the "other sources" weights to straggle upward -- noticeable but far from offering any basis for explaining their difference from the air values.
- a possible greater variability of the "other sources" weights, over and beyond this straggling.

Here schematic plots have done everything that we could ask.

(* easing the task *)

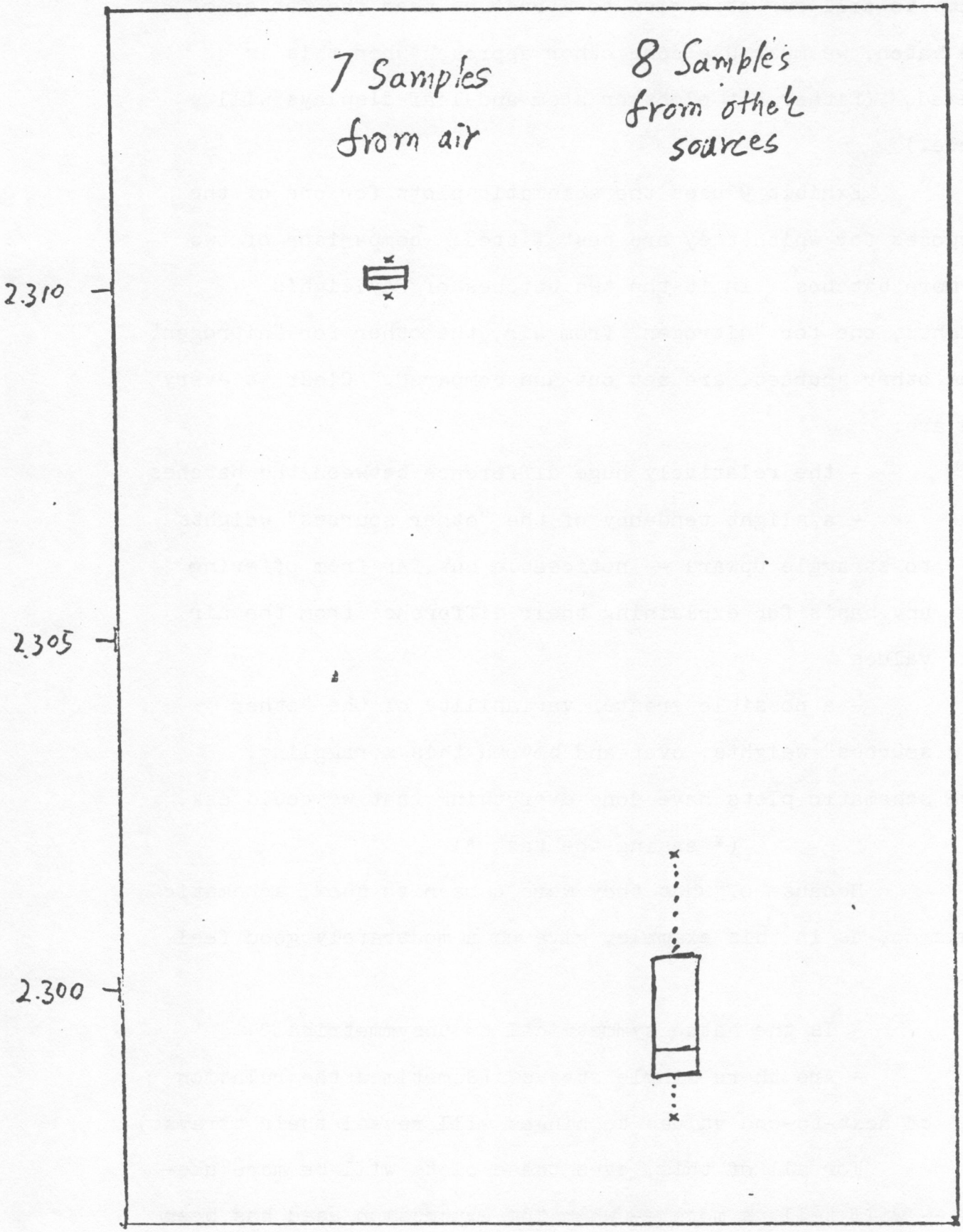
Because of what they were chosen to show, schematic plots do, as in this example, give us a moderately good feel for:

- Is the batch symmetrical or unsymmetrical?
- Are there single strays? (Sometimes the relation of next-to-end values to hinges will reveal their strays.)

For all of this, even these plots will be more useful -- will tell us more -- when the expression used has been chosen to approach the goals of:

exhibit 9 of chapter 5

Rayleigh's two Batches of weights of "Nitrogen"



- a symmetrical batch.
- no unnecessary strays.

We ought not to place undue and unnecessary burdens on any visual presentation. When any kind of plot shows lack of symmetry, either of outer values or of the bulk of its values, we ought to consider alternative expressions.

The three plots of exhibit 5 not only show which expression is better, they show how much more of what is essential is shown by the third schematic plot, the one for root height. Once the loud signals about asymmetry are out of the way, we can look at what is left.

(* aim reiterated *)

The Rayleigh data lets us make an old point once more. Aims can differ, and plots should follow aims. We can try to show either general behavior or detail. We should almost always choose one or the other.

Many books on graphical presentation stress getting the proper impression of relative size across to the reader. This may be sound for most articles in the daily newspaper, where making small differences more visible by starting the vertical scale far from zero is usually thought of as a "journalistic" or "political" trick.

We need to remember that doing just this is a vital scientific tool. If we had started the vertical scale at zero, and used bar graphs, Rayleigh's results -- so clear in exhibit 5 -- would have appeared as in exhibit 10. Whatever

exhibit 10 of chapter 5

Rayleigh's weights again

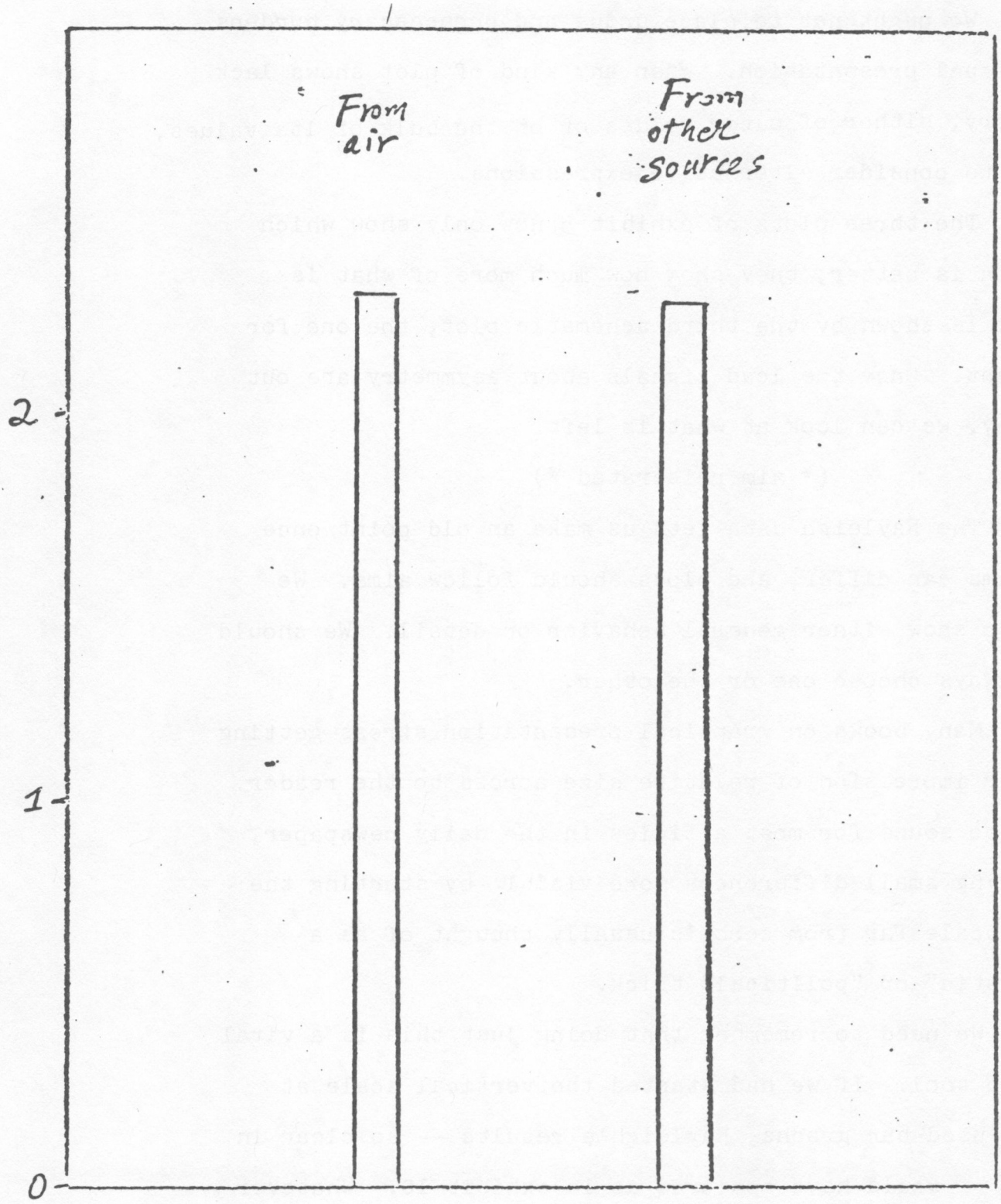


exhibit 10 may look like, it is neither evidence for a new chemical element nor a basis for a Nobel prize.

From the point of view of freight charges for compressed gases, exhibit 10's impression of "negligible difference" gives a useful impression, but it takes exhibit 9 to make clear the scientific issue. There is often no substitute for the detective's microscope -- or for the enlarging graph.

(* choice of plots *)

We are now well equipped. Schematic plots will be the standard form. But we will feel free to use box-and-dots instead whenever we feel that this will help. (There may even be situations where the simple box plot is worth using.)

Clearly we can use schematic plots in comparing three, four, or any reasonable number of batches. (When the number of batches becomes large, we will probably want to squeeze our bars down to solid lines, and thin our dotted lines out.)

As a visual approach to most comparisons, especially when supplemented by more detailed looks at the residuals, these plots work very well, often providing very effective comparisons.

Comparing more than a few is something that stem-and-leaf cannot do at all well. A display involving more than two -- or perhaps three -- stem-and-leaf displays tends to leave us with no clear impression.

The key advantages of schematic plots are, then:

- focussing attention where it is most likely to be useful.

- reducing "busyness" to a point where many batches can be compared.

5F. Some further examples

It is time to look at some further examples. Let us begin with the comparison of two moderately large batches.

(* ultimate hydroelectric powers *)

Exhibits 1 and 4 of chapter 3 set out the logs of the ultimate power in megawatts of two batches of hydroelectric power plants, one batch operated by the Bureau of Reclamation, the other by the Corps of Engineers. The compressed notations are, respectively:

[34, *HMH*: 1.18, **1.52**, 1.86, **2.21**, 3.30]

[62, *HMH*: 1.15, **1.83**, 2.04, **2.45**, 3.44]

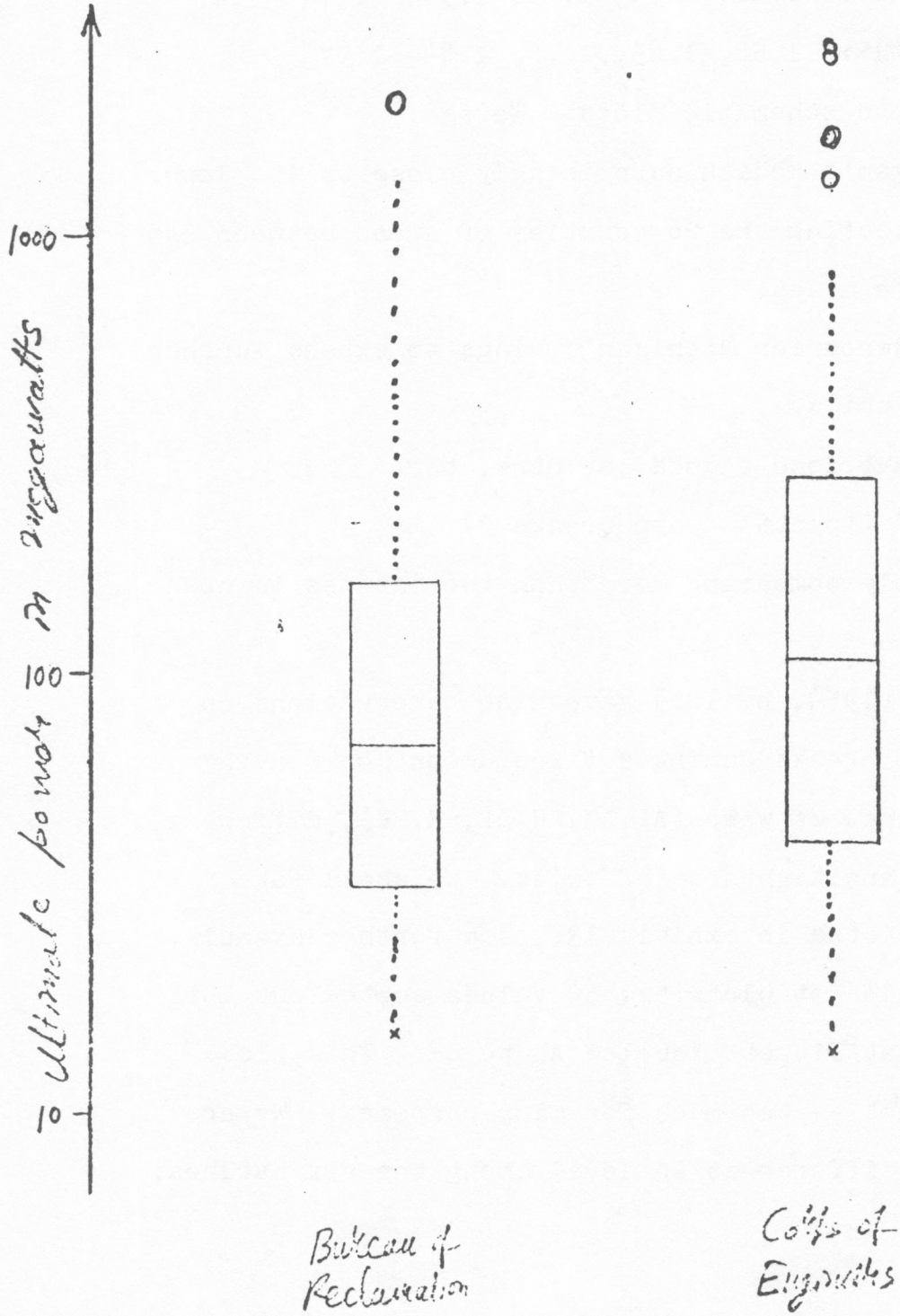
Exhibit **11** shows the corresponding schematic plots. (The Bureau of Reclamation batch is cut off at the level of the lowest point given; the extent to which the Corps of Engineers batch is cut off is unknown.) In this visual summary we see:

- a clear tendency for the power plants of the Corps of Engineers to be somewhat larger.

- a tendency of both batches to stray further toward higher log powers than toward lower log powers.

Schematic plots have clearly done well here.

Sizes of hydroelectric power plants compared
(see exhibits 1 and 4 of chapter 3)



* land areas of counties *

Exhibit 12 of chapter 3 showed land areas -- logs of areas in units of 10 square miles -- for the counties of Michigan and Mississippi. The condensed notations are, respectively:

[83, *HMH*: 1.51, **1.74**, 1.76, **1.94**, 2.17]

[82, *HMH*: 1.58, **1.65**, 1.76, **1.84**, 1.97]

Exhibit **12** shows the schematic plots. We see:

- Michigan's median surprisingly close to its lower hinge -- reflecting the 26 counties of areas between 545 and 595 square miles.

- a tendency for Michigan's wings to extend further than Mississippi's.

Schematic plots have done a good job here, too.

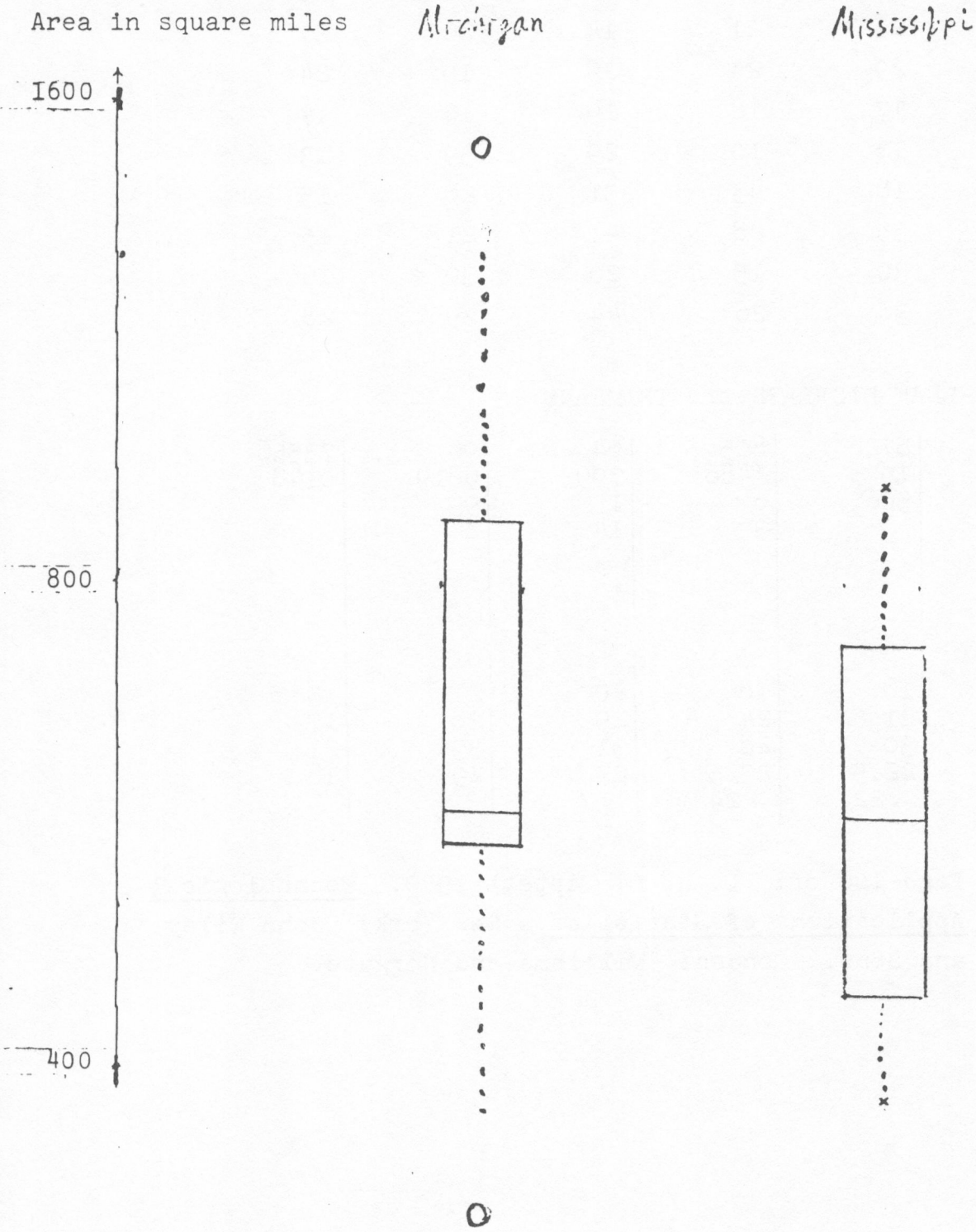
(* Tippett's warp breaks *)

An example comparing more than two batches is now in order.

Tippett [1950, p. 106] gave nine observations on the number of warp breaks during a fixed amount of weaving for each of six types of warp (AL,AM,AH,BL,BM,BH), differing in kind of cotton and tightness of twist. We shall take these 54 counts, listed in exhibit **13**, as a further example.

Exhibit **14** dot plots the 54 values sorted out into the six batches that Tippett labeled AL to BH. This plot shows us quite a lot -- too much for many purposes. We are given ideas about differences in level among the six batches;

Land areas of counties in states with
the same median area
(see exhibit 12 of chapter 3)



The 54 counts of warp breaks

A) DATA

AL	AM	AH	BL	BM	BH
26	18	36	27	42	20
30	21	21	14	26	21
54	29	24	29	19	24
25	17	18	19	16	17
70	12	10	29	39	13
52	18	43	31	28	15
51	35	28	41	21	15
26	30	15	20	39	16
67	36	26	44	29	28

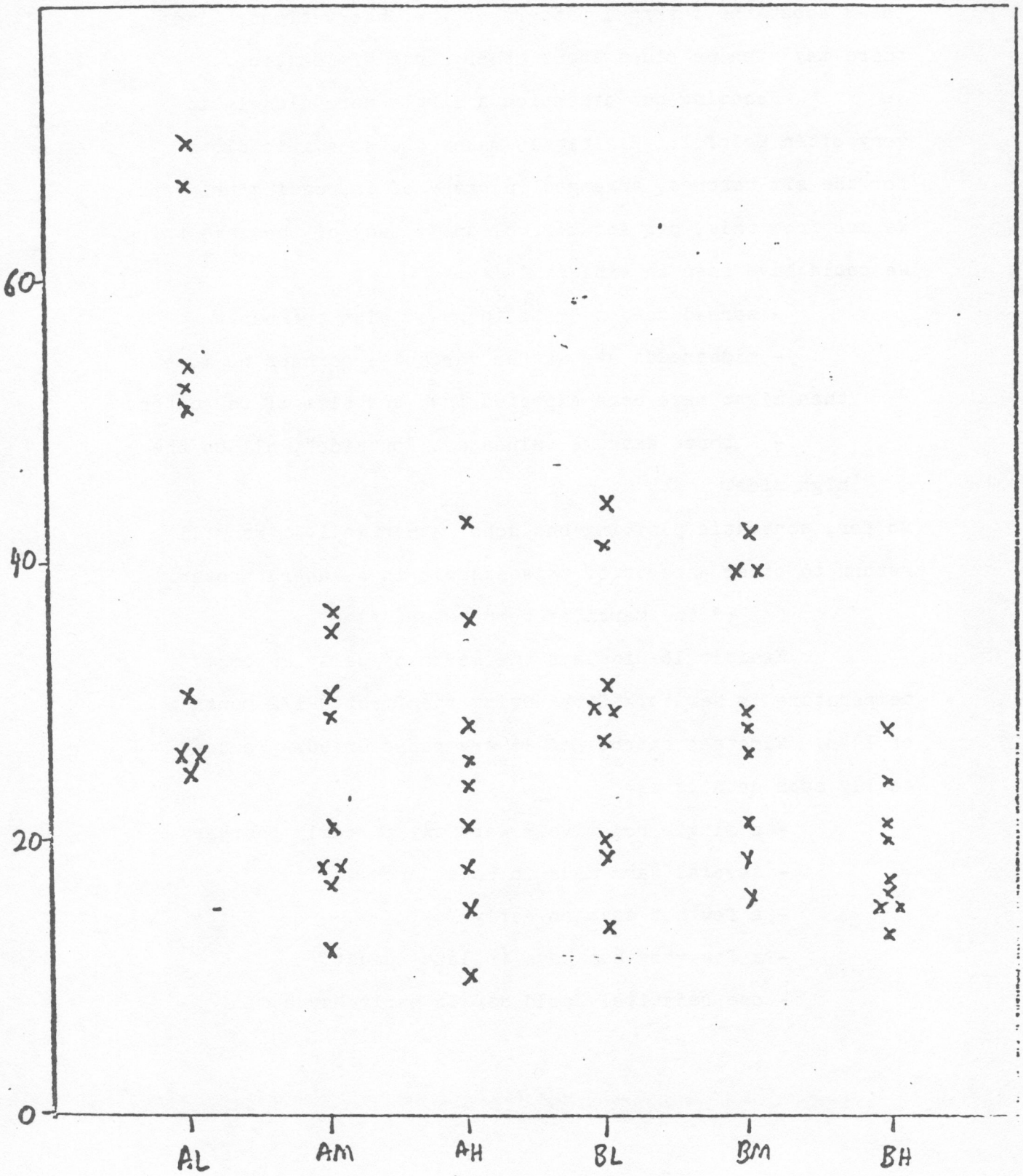
B) STEM-and-LEAF DISPLAYS and TRIMEANS

1		8728	805	94	96	73556
2	656	19	1486	7993	6819	0148
3	0	506	6	1	99	
4			3	14	2	
5	421					
6	7					
7	0					
9						
3H	26	18	18	20	21	15
5M	51	21	24	29	28	17
3H	54	30	28	31	39	21
tri	45.5	22.5	23.5	27	29	18
	→ 46	→ 22	→ 24			

C) SOURCE: Page 106 of: L. H. C. Tippett 1950. Technological Applications of Statistics. New York: John Wiley and Sons. London: Williams and Norgate.

exhibit 14 of chapter 5

Tippett's data on number of warp breaks



AL rather clearly runs high, while BH runs low. We are given ideas about the spreads of the various batches of nine values, AL appearing broader and BH narrower. We are even given ideas about the internal pattern of the observations. And there may even be clues about other sorts of behavior.

Focusing our attention a little more closely is very often helpful. Exhibit 15 makes the schematic plots for the six batches, arranged in order of increasing trimeans. We see from this, perhaps more clearly, many of the same things we could have seen in exhibit 14:

- spread does seem to increase with trimean.
- midspreads are rather variable, perhaps no more than might have been expected from the size of the batch.
- three extreme values are "outside", all on the high side.

So far, schematic plotting has done rather well. (We will return to other aspects of this example in a later chapter.)

(* the march of temperature *)

Exhibit 16 displays the march of daily maximum temperature in New York City during the first 9-1/2 months of 1948. Nineteen batches of 15 are represented. We see easily such details as:

- a single relatively warm day in early February.
- several warm days in late March.
- a few hot days in early May.
- a few very hot days in late August.
- one relatively cold day in early March.

Tippett's warp breaks in schematic plots

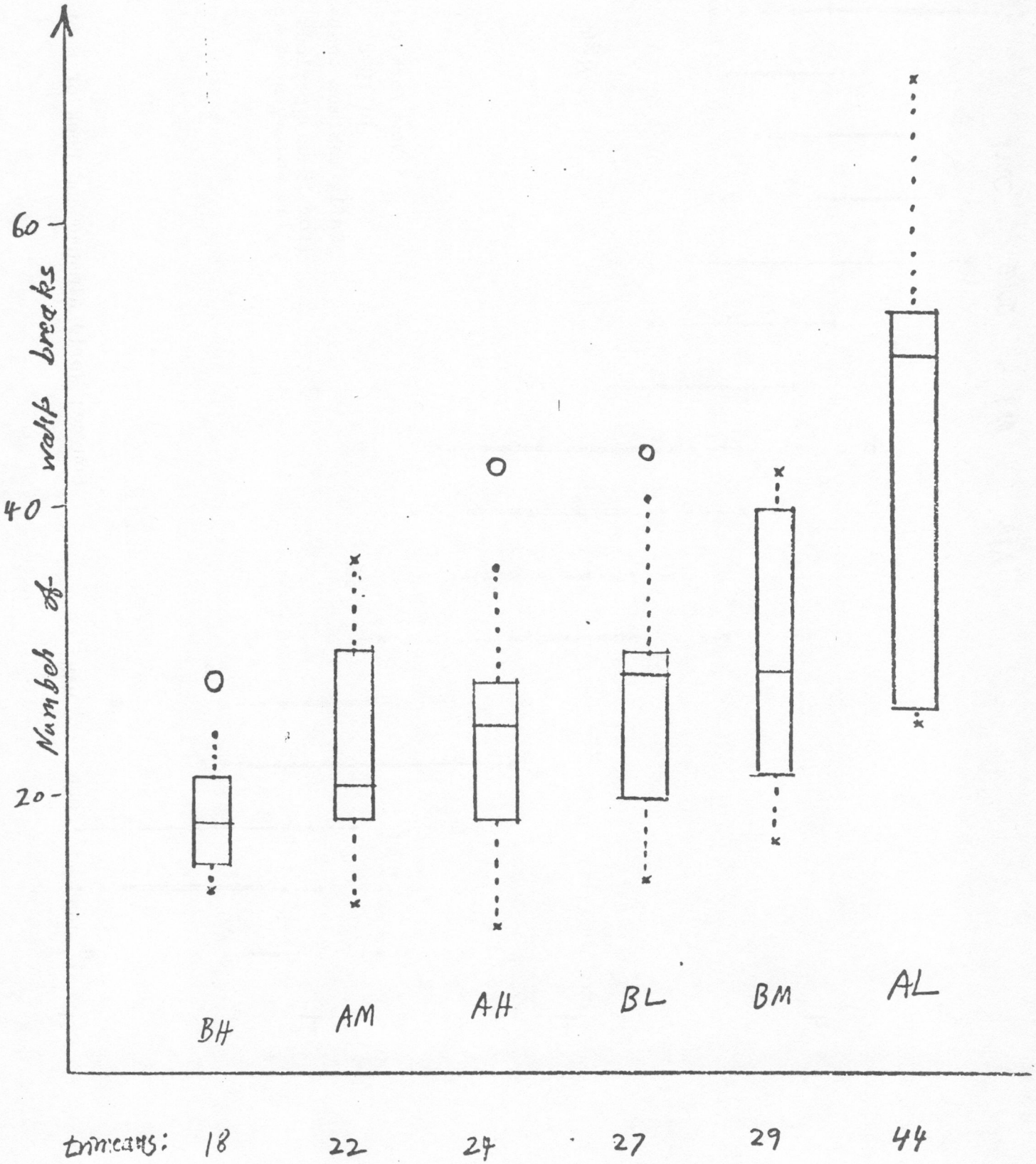
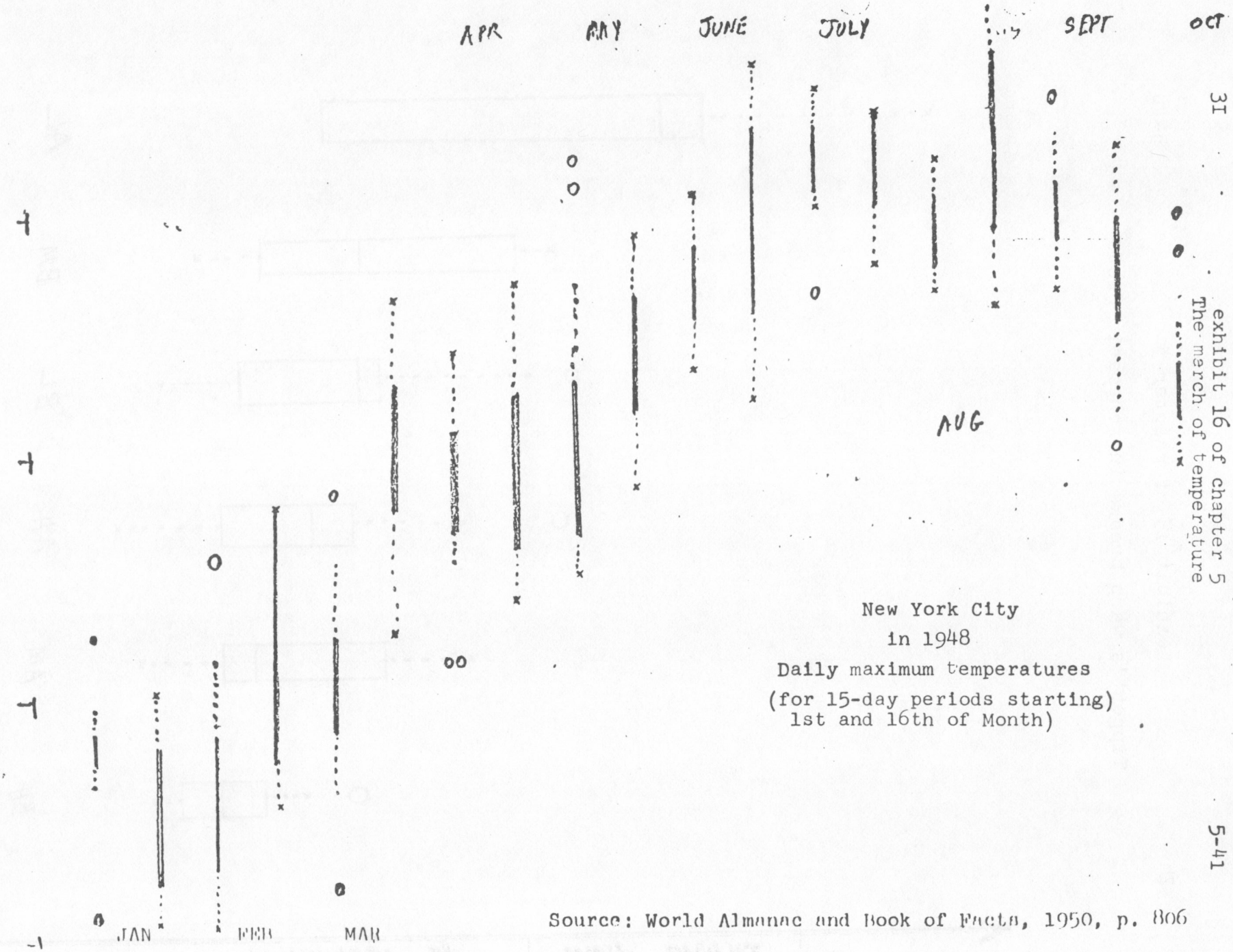


exhibit 16 of chapter 5
The march of temperature



New York City
in 1948
Daily maximum temperatures
(for 15-day periods starting)
1st and 16th of Month)

Source: World Almanac and Book of Facts, 1950, p. 806

- one relatively cool day in late June.
- two relatively cool days in late September.

However, these details do not interfere with broader insight, such as:

- the broad swing through the year, low near January-February and high in July.
- unusual warmth for late March as a whole (and for most of late August).

In this example, schematic plots have served us well. (No other technique we have considered would have been even barely satisfactory.)

It is nice to understand daily maximum temperatures. It would be nice to understand daily minimum temperatures at the same time.

In all, 5 temperatures were detached, 3 high and 2 low, where anything round about 1-4 would cause no remark. Similarly, 9 temperatures were outside but not detached, where round about 14 would cause no remark. Clearly, no remark is called for.

5G. Different aims, different schematic plots

We have developed a schematic plot that does a good job of

giving 5-number summaries schematically.

This is not the only kind of aim that a schematic plot might have. It is time to look into other useful aims.

What kinds of information does the kind of schematic plot we have met provide? Mainly these three:

- a central value (the median).
- an indicator of unstraggling spread (the difference of hinges.)
- indicators of straggling (the extremes, and the other individually given points).

We are usually happy with all but the first. Between the two of them, the midspread (difference of hinges) and the range (difference of extremes) give us a reasonable hold on spread. The individually given points ordinarily tell us what we need to know about straggling.

While we know how to locate the trimean from the barred box, there are situations where we would like to have the trimean given explicitly, rather than the median. We do this by using a dot for the trimean -- and omitting the bar for the median.

Exhibit 17 shows the two forms of schematic plot for Rayleigh's 15 weights of nitrogen. Exhibit 18 compares the land areas of counties in Michigan and Mississippi using schematic plots with trimeans. Exhibit 19 shows the same with a dotted box. If we want to make the comparison of the two batches, exhibit 18 does better than exhibit 12. If we want our attention called to the many Michigan counties of about the same size, this is done better by exhibit 12. Exhibit 19 focusses more attention on the trimean, but may not be an improvement on exhibit 18. Different aims call for different plots.

If we have several or many batches to compare, the trimean schematic plot continues to work well. Exhibit 20 gives this kind of plot for Tippett's warp breaks -- exhibit 21 does them with dashed boxes. With as many batches as this, the dotted boxes do better. Exhibit 22 plots the New York City march of temperature again -- notice how well bulging dots fit on line segments in this case.

exhibit 17 of chapter 5

Two forms of schematic plot for

Rayleigh's 15 weights of nitrogen

(see exhibit 8)

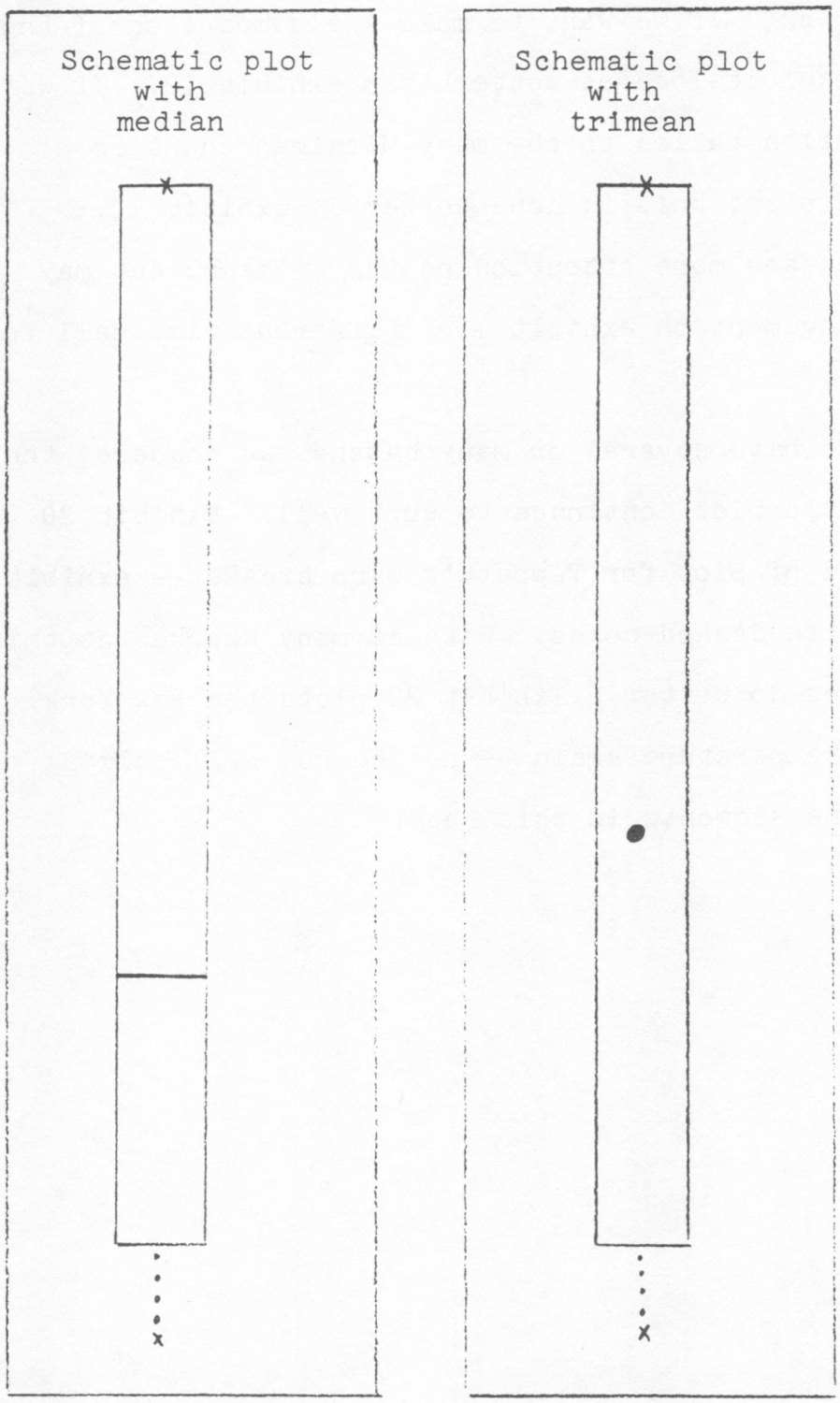


exhibit 18 of chapter 5

Land areas of counties in two states

(compare exhibit 12)

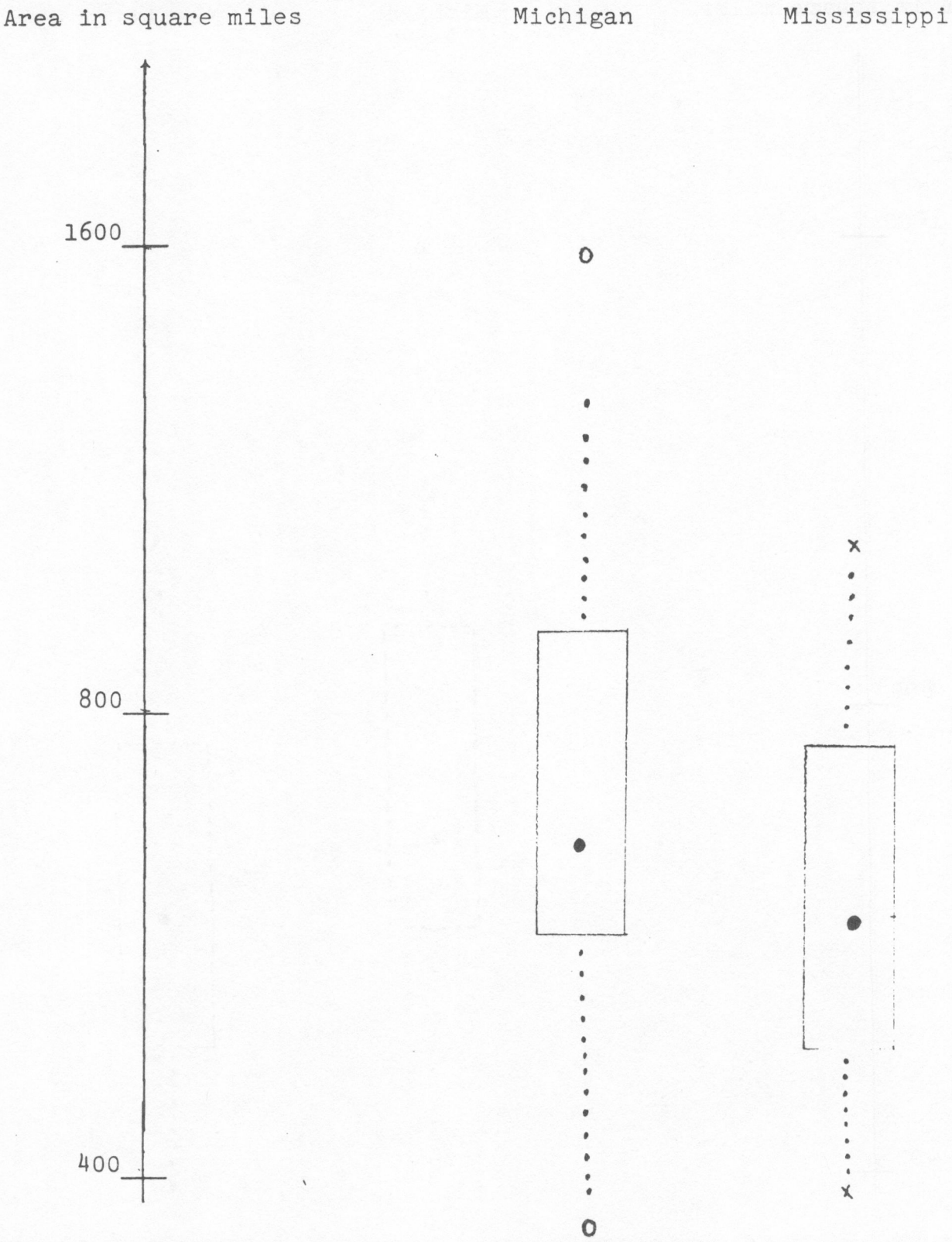
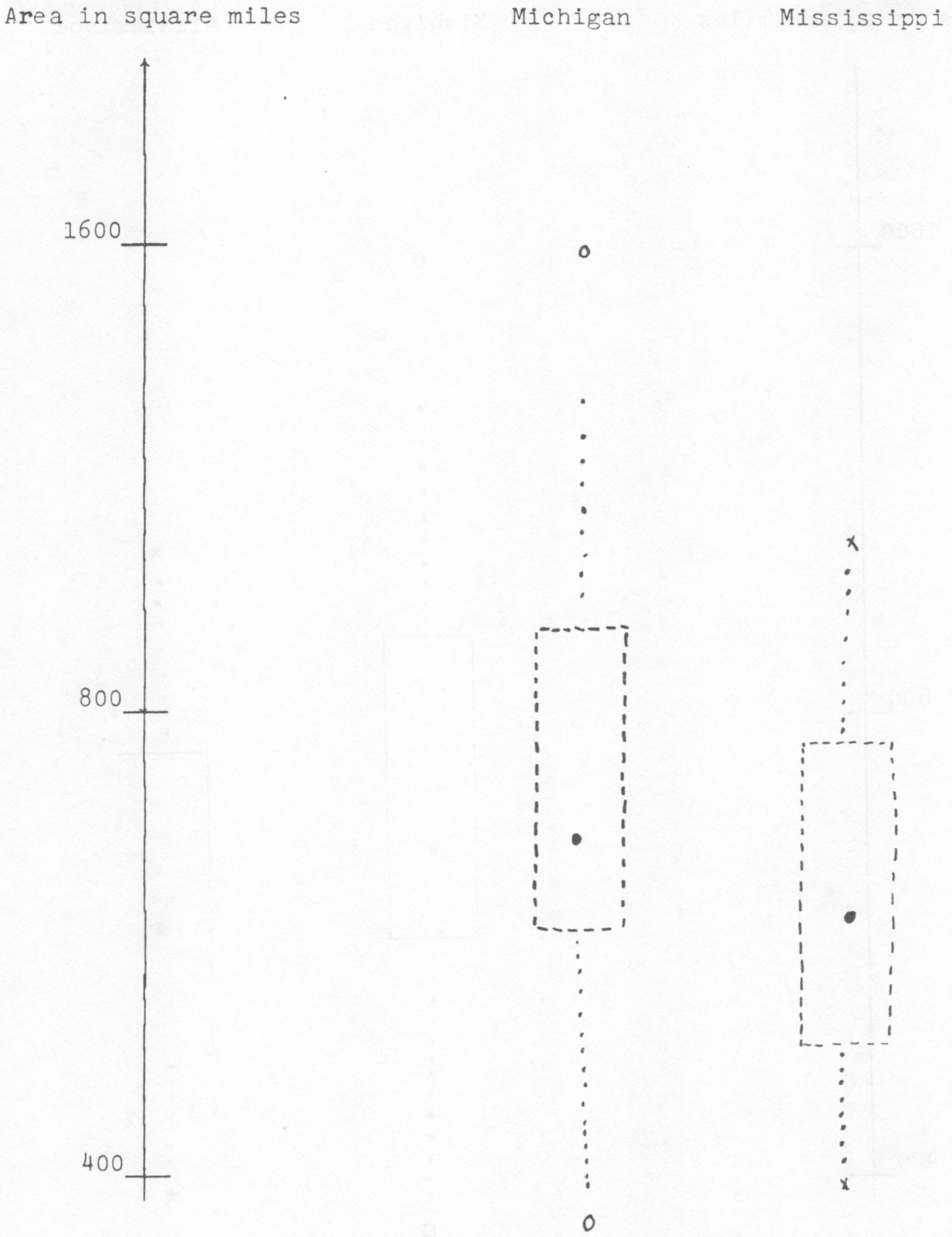


exhibit 19 of chapter 5

Land areas of counties in two states

(compare exhibits 12 and 18)



Tippett's warp breaks again
(compare exhibit 15)

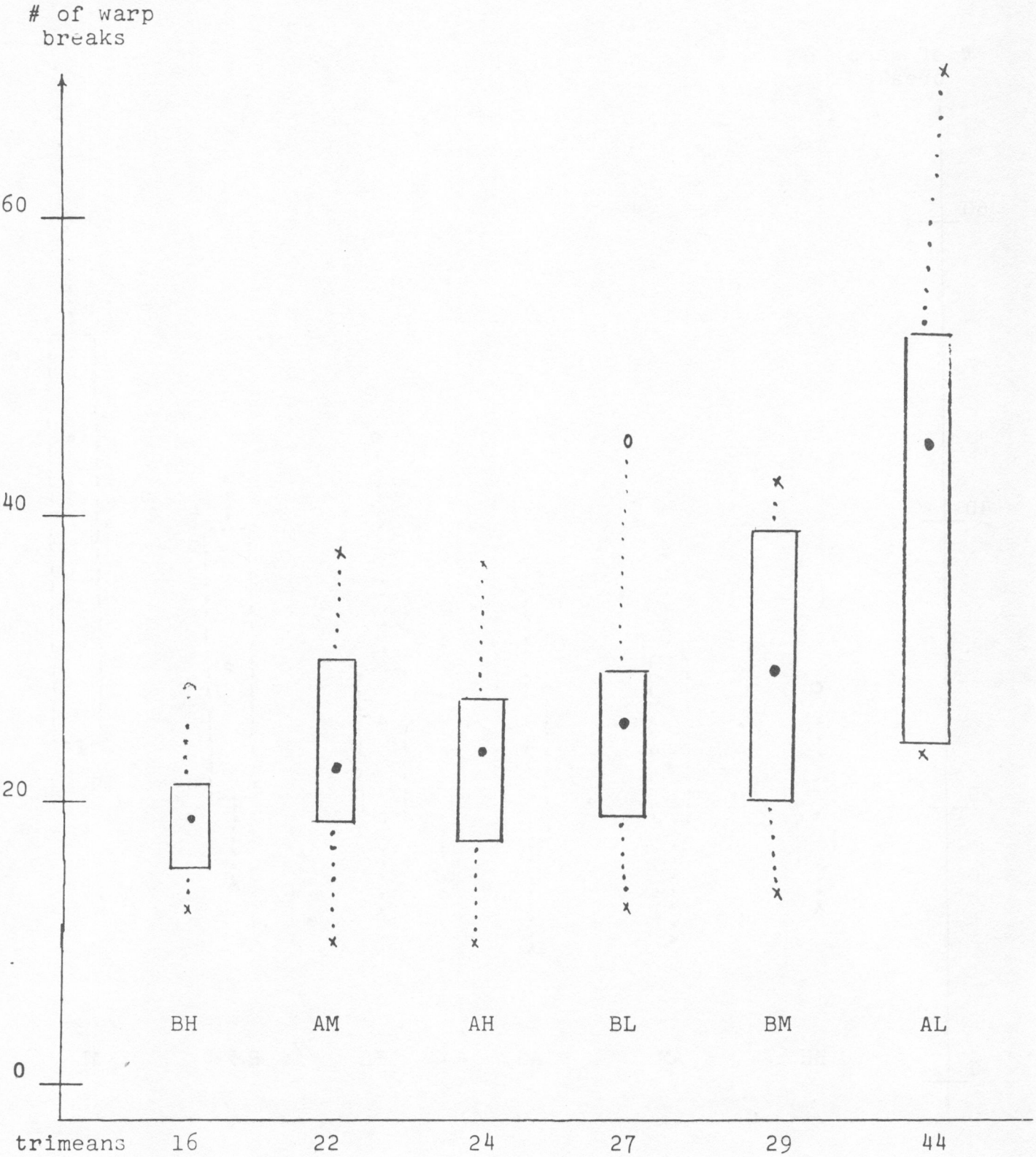


exhibit 21 of chapter 5

Tippett's warp breaks yet again
(compare exhibits 15 and 20)

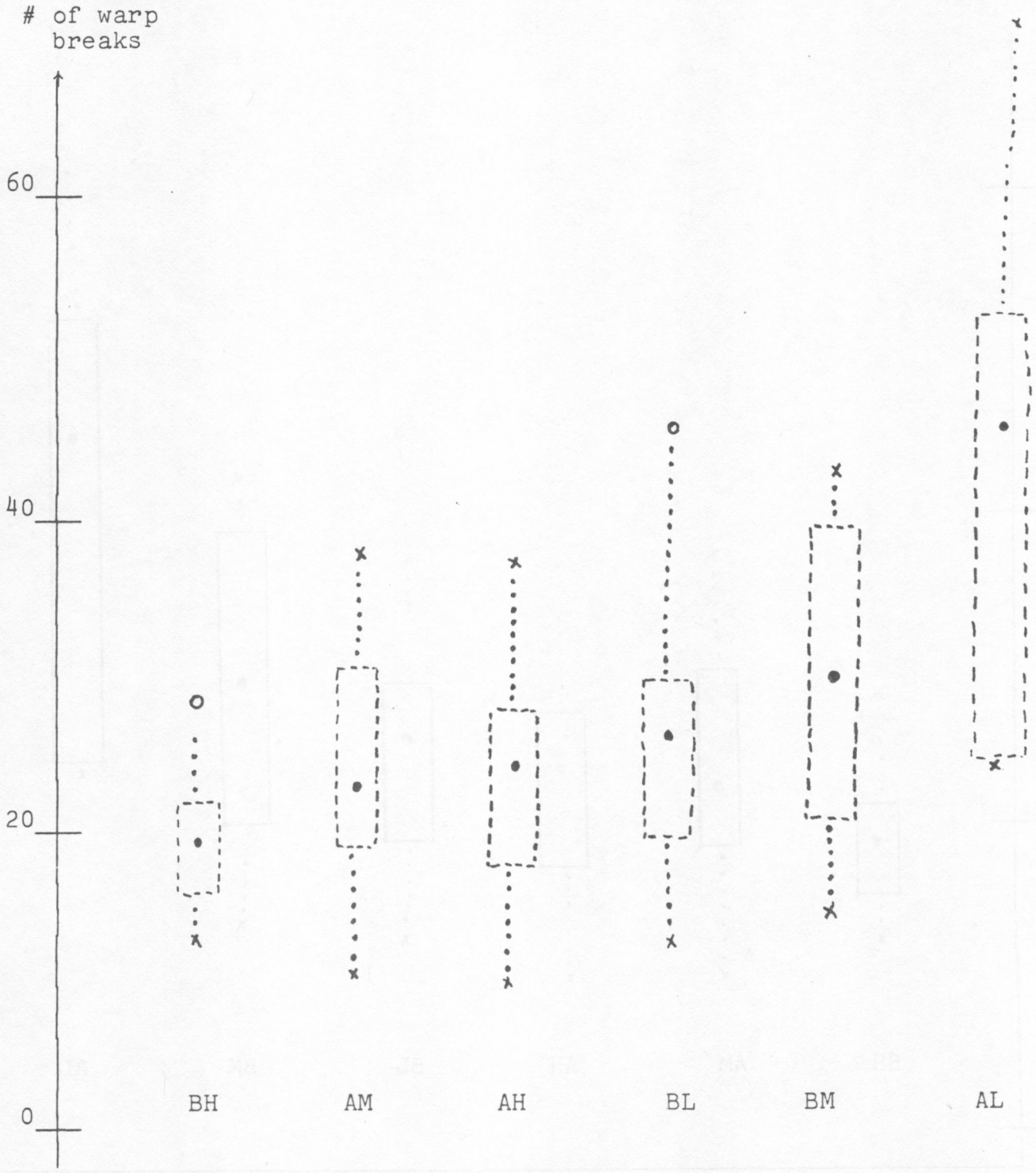


exhibit 22 of chapter 5

The march of temperatures again

(compare exhibit 16)

*****

5H. How far have we come?

In this chapter we have concentrated on summarizing batches visually.

Plots of the individual values of the batch have been shown to emphasize either centering or spread and pattern, depending on how we choose to draw them. Dot plots have the free-swinging graph's great advantage of showing us the unexpected, but they do sometimes seem to divert our attention from the simple and always give a poor return for hand labor when applied to large batches.

Box plots show us graphically what the five-number summaries tried to convey numerically. They do summarize, they are visually effective, but we tend to seek a little more detail. If we wish only to emphasize centering, they are quite effective. In smaller batches, a combination of a (weakened) box plot with a detailed dot plot is effective in showing us spread and pattern.

Schematic plots are an effective means of representing and comparing those aspects of batches that are most likely to be useful, when, as is usual, interior details are not important.

By showing both detached and outside observations individually, and the remaining observations collectively, they show us things we are most likely to need to see. They can do this for many batches at the same time. To get better impact where we need it most often, we must, of course, give

up something. For schematic plots what we give up is mainly detail , but we can miss major details in the middle of the batch.

Outside observations -- shown as solid circles -- are those beyond the side values, each of which is one difference of hinges outside the corresponding hinge .

Detached observations -- shown as open circles -- are those beyond the three-halves values, each of which is 3/2 (difference of hinges) outside the corresponding hinge.

Dotted lines extend out from the barred box to the adjacent observations -- the outermost observations not beyond the side values.

Barred boxes show hinge, median, and hinge.

Dotted boxes show hinge, trimean, and hinge. The use of dotted boxes often helps, especially when several batches are being compared.

******

5P. Problems for chapter 5

1. Make schematic plots for the batches of viscosines in exhibit 12 of chapter 1.

2. Do the same with the radial velocities in exhibit 13 of chapter 1.

3. Do the same with the association constants in exhibit 14 of chapter 1.

4. Do the same with 3 or more of the batches of atomic weights in exhibit 7 of chapter 2.

5. Do the same with 2 or more of the batches of gold contents in exhibit 8 of chapter 2.

6. Do the same with the tonnages in exhibit 24 of chapter 3.

7. Do the same with the areas of lakes in exhibit 25 of chapter 3.

8. Do the same with the areas of oceans and lakes in exhibit 26 in chapter 3.

9. Do the same for the river lengths in exhibit 27 in chapter 3.

10. Do the same for the reservoir capacities in exhibit 29 in chapter 3.

11. Do the same for the two batches of New York city snowfalls in problems (1) and (2) of section 4P.

12. Do the same for the batch of carobene percentages.

13. Do the same for the fabric drapes of exhibit 18 of chapter 4.

14. Do the same for at least 3 of the samples of carbon monoxide determinations in exhibit 19 of chapter 4.

15. Do the same for the nitrogen ultramicrodeterminations in exhibit 22 of chapter 4.

16. Do the same for at least one size of termites in exhibit 24 of chapter 4.

17. Do the same for the batches of critical temperatures in exhibit 23.

exhibit 23 of chapter 5

Critical temperatures in centigrade for various substances

A) The DATA

Ethyl ether (page 254)	Benzene (page 256)	Hydrochloric acid (page 257)	Carbon dioxide (page 258)	Oxygen (page 261)	Nitrogen (page 262)
190.0	280.6	52.3	31.9	-118.0	-146.0
195.5	291.7	51.50	31.35	-118.8	-146.0
191.8	296.4	52.2	31.4	-118.0	-146.5
193.7	290.5	51.8	30.98	-118.82	-147.13
197.0	288.1	51.0	31.10	-118.0	-147.13
194.7	288.68	51.4	31.6		-147.7
193.6	288.5	51.45			
194.4					
193.6					
193.04					
193.8					
193.26					

B) The SOURCE

Landolt-Bornstein (5th) 1923.